Project 4: Washington and Lee ChatGPT using Phi2 and QLora

Overview

This assignment focuses on creating a ChatGPT-like conversational AI model specifically tailored for Washington and Lee University faculty, staff, and students. You will leverage Microsoft's Phi2 language model and fine-tune it using the QLora technique with documents from W&L's website or internal documents from various campus departments. The goal is to create a knowledgeable and helpful AI assistant that can answer questions and provide information relevant to the W&L community.

The assignment consists of several key components. First, you will gather and preprocess a dataset of W&L-related documents to be used for fine-tuning the Phi2 model. Next, you will apply the QLora technique to efficiently adapt the model to the W&L-specific knowledge. Finally, you will develop a user-friendly Gradio app that serves as the interface for interacting with the trained model via an API.

Implementation Steps

1. Data Preparation

   a. Gather relevant documents from W&L's website and various campus departments.

   b. Preprocess the collected documents to ensure compatibility with the QLora fine-tuning process.

2. Fine-tuning with QLora

   a. Set up the necessary environment and dependencies for using Phi2 and QLora.

   b. Fine-tune the Phi2 model using the preprocessed W&L documents and the QLora technique.

   c. Evaluate the fine-tuned model's performance on a held-out validation set.

3. Gradio App Development

   a. Design and implement a Gradio app that provides an intuitive interface for users to interact with the fine-tuned model.

   b. Integrate an API to facilitate communication between the Gradio app and the trained model.

   c. Ensure the app is responsive, user-friendly, and visually appealing.

4. Documentation and Reporting

a. Prepare a report detailing the steps involved in data preparation, model fine-tuning, and app development.

b. Document any challenges encountered and how they were addressed.

c. Provide instructions for setting up and running the Gradio app.

Turn In

- The preprocessed W&L document dataset used for fine-tuning.

- The code for fine-tuning the Phi2 model using QLora.

- The Gradio app code and associated files.

- A report covering the implementation process, challenges faced, and instructions for running the app.

- A live demo of the Gradio app showcasing its functionality and performance.

Note: Ensure that the fine-tuned model and the Gradio app adhere to ethical guidelines and do not generate inappropriate or offensive content. The AI assistant should provide helpful and relevant information to the W&L community while maintaining a respectful and inclusive tone.

Example: https://github.com/brevdev/notebooks/blob/main/phi2-finetune-own-data.ipynb